Effect of metal pollution on genetic variation in natural populations of selected soil invertebrate species with different dispersal potential

Giska, I.

2016

document version
Publisher's PDF, also known as Version of record

Chapter 4

Genome-wide genetic diversity of rove beetle populations along a metal pollution gradient

Iwona Giska, Wiesław Babik, Cornelis A.M. van Gestel, Nico M. van Straalen, and Ryszard Laskowski

Published: Ecotoxicology and Environmental Safety 2015, 119:98-105

Abstract

To what extent chemical contamination affects genetic diversity of wild populations remains an open question in ecotoxicology. Here we used a genome-wide approach (615 nuclear RADseq loci containing 3017 SNPs) and a mtDNA fragment (ATP6) to analyze the effect of long-term exposure to elevated concentrations of metals (Cd, Pb, Zn) on genetic diversity in rove beetle (*Staphylinus erythropterus*) populations living along a pollution gradient in Poland. In total, 96 individuals collected from six sites at increasing distance from the source of pollution were analyzed. We found weak differentiation between populations suggesting extensive gene flow. The highest genetic diversity was observed in a population inhabiting the polluted site with the highest metal availability. This may suggest increased mutation rates, possibly in relation to elevated oxidative stress levels. The polluted site could also act as an ecological sink receiving numerous migrants from neighboring populations. Despite higher genetic diversity at the most polluted site, there was no correlation between the genetic diversity and metal pollution or other soil properties. We did not find a clear genomic signature of local adaptation to metal pollution. Like in some other cases of metal tolerance in soil invertebrates, high mobility may counteract possible effects of local selective forces associated with soil pollution.

Keywords

Staphylinus erythropterus, genetic variation, gene flow, heavy metals, RADseq, mtDNA

Introduction

Anthropogenic activities lead to a range of negative changes in the environment, including chemical contamination. This constitutes a challenge for populations to survive unfavorable conditions. Large healthy populations with high levels of genetic diversity can effectively cope with novel selection pressures (Frankham et al. 2010). However, chronic exposure to pollution may decrease genetic diversity of populations diminishing their evolutionary potential (Van Straalen and Timmermans 2002; Dallinger and Höckner 2013).

Population-level effects start, however, in individual organisms exposed to pollution. Toxic chemicals cause damage to their body cells and disturb their physiology. It has been suggested that organisms exposed to toxicants face a trade-off associated with energy allocation to competing metabolic processes, including detoxification (Sibly and Calow 1989; Posthuma and Van Straalen 1993). Therefore, life history traits such as growth, reproduction and survival are often affected due to increased metabolic expenditure to detoxification (Jones and Hopkin 1998; Spurgeon et al. 2000). This may result in reduced individual fitness and decreased size of wild populations. In addition, when populations have become small, genetic drift may remove variation faster (Gillespie 1998). Eventually a population may enter the "extinction-vortex" (Frankham et al. 2010).

The population genetic responses to pollution were recently grouped into four categories called "the four cornerstones of Evolutionary Toxicology" (Bickham 2011). These include: genome-wide changes of genetic diversity, changes in allele frequency distribution due to selection, changes in population differentiation due to altered dispersal patterns, and changes in allele frequency due to increased mutation rate. Clearly, the key feature in all these categories is genetic diversity. Pollution impacts population genetic diversity by affecting four major evolutionary processes: genetic drift, selection, migration, and mutation.

Results of the existing studies aiming at assessing effects of metal pollution on genetic diversity of populations are inconclusive and contradictory. For example, reduction of genetic diversity was reported in populations of the marsh frog (*Rana ridibunda*) from contaminated wetlands of Sumgayit, Azerbaijan (Matson et al. 2006) and in sandhopper (*Talitrus saltator*) living at the metal-polluted Tyrrhenian coast in central Italy (Ungherese et al. 2010). Based on gene flow estimations,

Matson et al. (2006) concluded that the polluted Sumgayit region acted as an ecological sink characterized by high frequency of immigrant frogs from surrounding areas. On the other hand, Eeva et al. (2006), studying two bird species exposed to metals and nuclear radiation in Finland and Russia, reported decreased genetic diversity in populations of the pied flycatcher (*Ficedula hypoleuca*), but increased genetic diversity in the great tit (*Parus major*) living at polluted sites. The difference in species response was explained by different dispersal patterns and detoxification abilities. The genetic erosion hypothesis of Van Straalen and Timmermans (2002) was supported by the study of Andre et al. (2010). They investigated highly differentiated populations of the earthworm *Lumbricus rubellus* from a Pb-polluted habitat near Cwmystwyth, Wales, UK, and hypothesized a reduction of genetic diversity through the loss of distinct mtDNA genetic lineages. At the same time, a number of authors reported no influence of metal pollution on population genetic diversity in different animal species, for example the wood mouse *Apodemus sylvaticus* (Berckmoes et al. 2005), the earthworm *Dendrobaena octaedra* (Simonsen et al. 2008) or the ground beetle *Pterostichus oblongopunctatus* (Lagisz et al. 2010). Therefore, the question to what extent long-term exposure to chemical pollutants affects genetic diversity of wild populations remains open.

Detecting genetic responses to environmental change is not straightforward in the absence of a priori candidate loci with a large effect on the phenotype of interest (Hoffmann and Willi 2008). In the above mentioned studies, researchers usually used small numbers of molecular markers. These were mainly single mtDNA genes, microsatellites, RAPD or AFLP. What seems to be lacking is a genome-wide approach which could allow for better insight into population genetic processes occurring across a whole genome. This was highlighted by Bickham et al. (2000) in their review of genetic effects of pollution in natural populations with the statement: “The challenge for the future lies in gaining an integrated perspective of the genomic response to contaminant exposure and to the forces that promote population divergence”. Recent developments of next-generation sequencing (NGS) technologies (Mardis 2013), applicable also for non-model species (Baird et al. 2008; Catchen et al. 2011; Peterson et al. 2012; Zieliński et al. 2014), and their decreasing costs, increase the potential for detailed population genomics research.

Here, we aimed at testing the association between population genetic diversity and soil contamination. We took advantage of next-generation sequencing methods to assess the impact of multigenerational exposure to elevated concentrations of metals (Cd, Pb, Zn) on rove beetle (*Staphylinus erythropterus*) populations living along a metal pollution gradient in Southern Poland. Genetic diversity, gene flow and population structure were estimated for both the mitochondrial and the nuclear genome using Restriction-site Associated DNA Sequencing (RADSeq). Estimated population genetics parameters were related to metal concentrations in soil. Our prediction was that *S. erythropterus* populations would show a reduction of genetic diversity with increasing pollution due to toxicity, population size decline (bottleneck), and directional selection. We expected this effect to be dispersal dependent as higher dispersal potential results in more intensive gene flow which may overcome the selection and bottleneck impact. Here we report results concerning just the most dispersive species, however, in the whole project we test several species with different dispersal capabilities.

Materials and methods

Staphylinus erythropterus

The rove beetle *Staphylinus erythropterus* Linné, 1758, belongs to the family *Staphylinidae* – one of the largest families of beetles (*Coleoptera*) (Herman 2001). It is a common and widespread species, distributed in nearly all zoogeographical regions. It can be found throughout Poland, living mainly on the soil surface and in the litter layer in forests. As a holometabolous insect, it has a four-stage life cycle, including egg, larva, pupa and imago. There is only one generation of *S. erythropterus* during a year (Szujecki 1980). Eggs are laid in May and June, usually on the soil surface. To develop they must absorb water from the environment. The predatory larva lives in deeper soil layers. Metamorphosis of pupa takes place in August, in soil chambers dug by the larva (Szujecki 1980). Adults, also predators, feed on soft-bodied invertebrates (e.g. nematodes, mites, springtails, slugs, earthworms) and other insect larvae or pupae. They are able to fly actively and dispersal is most intensive in Spring during the reproduction season. As all life stages of *S. erythropterus* live in the top-soil litter layer, where most emitted metals end-up, they are exposed to pollution, especially

considering the fact they absorb soil porewater and feed on animals recognized as macroconcentrators of metals. According to the review of Bohac (1999) Staphylinid larvae are considered more sensitive to pesticide and radionuclide pollution than imagos. However Zvereva et al. (2003) concluded that beetle larvae are less sensitive to metal pollution than adults because they are able to get rid of accumulated metals during moulting.

Sampling

The study was conducted at six sites located in two distinct areas in Poland, approximately 180 km apart (Fig. 1). Five sites (OL1-OL7) were located in the mining and smelting area at increasing distances from the zinc-and-lead smelter 'Bolesław' near Olkusz in Southern Poland (50°16′-50°32′N, 19°29′-19°38′E). One site (PK) was established in a clean area in the Kozienicka Forest, Central Poland (51°29′34″N, 21°16′16″E). All sites were characterized by similar habitat type - mixed pine forest on acidic soils with well-developed mor organic layer.

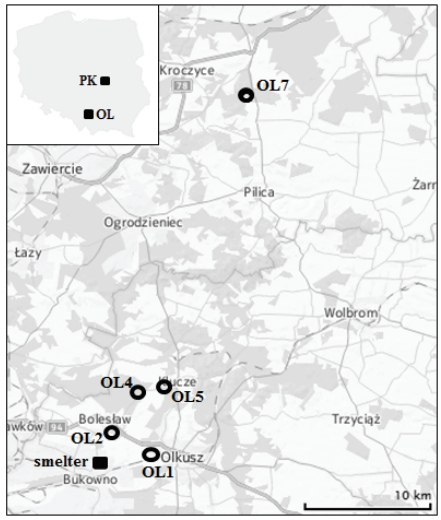


Fig. 1. Location of the two study areas in Poland (insert): PK – Kozienicka Forest, OL – Olkusz area, and detailed location of the Olkusz sites (OL1-OL7); gray patches represent forests. See text and Table 1 for sites characteristics.

Major metals contaminating the studied area were Cd, Pb and Zn, being present in a broad range of concentrations (Table 1). Detailed description of soil analysis procedures is available elsewhere (Giska

et al. 2014). Other metals (Co, Cr, Cu, Ni), although present, were usually found at rather low concentrations (Pasieczna and Lis 2008; Simonsen et al. 2008, Giska et al. 2014).

Table 1. Characteristics of the sites used for sampling the rove beetle *Staphylinus erythropterus*. Shown are distances from the 'Bolesław' smelter, soil pH in 0.01 M $CaCl_2$, organic matter content of the ~10 cm upper soil layer (OM%), soil metal concentrations [mg kg^{-1} dwt.]: total concentrations in normal font and 0.01 M $CaCl_2$-extractable concentrations in italics; mean ± SD (n=3). Part of the data shown here were taken from Giska et al. (2014).

Site	Distance[km]	pH_{CaCl2}	OM [%]	Cd [mg kg^{-1}]	Pb [mg kg^{-1}]	Zn [mg kg^{-1}]
OL1	3.3	5.06±0.06	45.1 ± 1.3	63.2 ± 3.0	3 041 ± 158	7 991 ± 536
				0.892 ± 0.018	*0.553 ± 0.014*	*54.4 ± 1.1*
OL2	2.5	4.12±0.03	53.5 ± 0.4	49.1 ± 1.1	2 060 ± 37	3 960 ± 54
				3.69 ± 0.05	*1.98 ± 0.05*	*211 ± 2*
OL4	5.3	3.46±0.02	54.2 ± 2.0	14.8 ± 0.2	847 ± 38	966 ± 22
				1.98 ± 0.02	*1.88 ± 0.02*	*84.3 ± 1.5*
OL5	7.7	4.29±0.01	36.3 ± 0.7	12.1 ± 0.7	708 ± 12	756 ± 11
				0.688 ± 0.012	*0.526 ± 0.015*	*30.8 ± 0.1*
OL7	~32	4.25±0.01	6.68 ± 0.08	1.36 ± 0.10	60.9 ± 3.5	88.6 ± 13.6
				0.393 ± 0.006	*0.231 ± 0.007*	*16.8 ± 0.7*
PK	~180	2.82±0.01	35.4 ± 1.3	0.677 ± 0.102	56.5 ± 3.9	36.5 ± 3.8
				0.198 ± 0.069	*0.376 ± 0.032*	*6.98 ± 1.39*

Adult rove beetles were collected alive using pitfall traps. The traps were installed at ~200 m^2 plots at each study site. From each site we collected 16 live individuals, 96 in total. They were washed with deionized water and starved individually in perforated plastic containers with moist filter paper for at least 24 h. Then, they were preserved in 96% ethanol and stored. Taxonomic identity of the species was confirmed with the identification key of Szujecki (1980). By inspecting the abdomen and presence of male reproductive organ (aedeagus) under a magnifying glass, the sex of each individual was determined. Genomic DNA was extracted from head and thorax tissues of ethanol-preserved specimens using Wizard® Genomic DNA Purification Kit (Promega). The purity and concentration of extracted genomic DNA was determined with NanoDrop spectrophotometer and Qubit® fluorometer. Sequence data for mitochondrial marker development were obtained by 454 pyrosequencing (1/16 of the PTP plate; GS FLX + Titanium (Roche)) of one randomly selected individual.

Mitochondrial ATP6 sequencing

The mitochondrial ATP6 (ATP synthase 6) gene sequence of *S. erythropterus* was reconstructed from 454 reads using blast searches against mitochondrial genomes of several beetle species. Primers for PCR amplification (F1/R1; Table S1.1) were designed with Primer3 software (Koressaar and Remm 2007; Untergrasser et al. 2012). PCR reactions were performed in 15 µl (~50-150 ng of DNA template, 0.5 µM of each primer, 1X *Taq* buffer with $(NH_4)_2SO_4$, 1.5 mM of $MgCl_2$, 0.2 mM of each dNTP, 0.75 U of *Taq* polymerase (Thermoscientific)) under cycling conditions described in Table S1.1. Products were visualized on an agarose gel. After Exo-AP cleaning (Exonuclease I and Thermosensitive Alkaline Phosphatase; Thermoscientific) PCR product was sequenced using the BigDye® Terminator v3.1 Cycle Sequencing Kit. Sequencing reactions were cleaned with Ethanol/EDTA precipitation method and run on the ABI 3130xl Genetic Analyzer (Applied Biosystems). Raw sequences were aligned and edited with SeqScape® software (Applied Biosystems).

In the case of some individuals with double peaks observed in the sequencing electropherogram, multiple additional approaches, as suggested by e.g., Sorenson and Quinn (1998) or Calvignac et al. (2011), were used to resolve whether it concerned heteroplasmy or nuclear mitochondrial pseudogenes (numts). These included i) digestion of PCR product with restriction enzyme (TseI; 5′-G^CWGC-3′) with the recognition site falling within one of the ambiguous sequence fragments, ii) DNA extracted from legs used for PCR as muscle tissue is enriched in mitochondria, iii) PCR with Phusion Hot Start II High Fidelity DNA polymerase to lower amplification error probability, iv) extreme dilutions of DNA (0.001 ng) used as PCR template to decrease the number of nuclear genome copies, and v) PCR with two more primers amplifying longer fragments (Table 2). After applying the above mentioned steps we still observed all the double peaks, and digestion with restriction enzyme confirmed the presence of two sequences. To avoid their effect on estimated population genetics parameters, we masked double peak positions with ‘N’ in all individuals. Because double peaks occurred only in a few positions, we argue that the effect of the possible numts or heteroplasmy on estimates of mtDNA diversity and differentiation among populations should be minor.

RADseq

Libraries for RAD sequencing were prepared according to double digest RADseq method described by Peterson et al. (2012). Briefly, for each individual 250 ng of genomic DNA was digested with SphI-HF and PstI-HF restriction enzymes (New England Biolabs). Adapters, one with 5 bp barcode, were then ligated using T4 DNA Ligase (NEB) in 40 µl reactions under the following conditions: $23^{o}C$ – 60 min, $65^{o}C$ – 10 min, cooling $0.022^{o}C\ s^{-1}$. After ligation, individual samples from each population were equally pooled resulting in six libraries. Purified libraries were then size selected with LabChip XT (LabChip XT DNA 300 Assay Kit; PerkinElmer). We aimed at selecting the 346 – 406 bp fraction, which according to Bioanalyzer analysis of digested genomic DNA and trial Illumina MiSeq sequencing, was supposed to result in no more than ~10,000 RAD tags per individual beetle. After size selection, libraries were amplified in PCR reactions (20 µl) containing: 1X Phusion buffer, 200 µM of each dNTP, 0.75 µM of PCR1 and PCR2 primer, 0.5 U of Phusion HF polymerase (Thermoscientific) and 2.5 µl of size selected library. The PCR profile was as follows: $98^{o}C$ –30 s, 12 cycles: $98^{o}C$ – 10 s, $62^{o}C$ – 30 s, $72^{o}C$ – 30 s and final extension $72^{o}C$ – 5 min. One of PCR primers included 6 bp index added to distinguish libraries of different populations. Amplified samples were inspected on Bioanalyzer (HS DNA chips; Agilent Technologies) to check their size distribution. Then, all six indexed libraries were pooled based on Qubit® measurements, and sent for Illumina HiSeq 2000 sequencing (single end, 100 bp) at the Center for Genome Research and Biocomputing, Oregon State University, USA (see Supporting information S1.2). Prior to sequencing molarity of the sample was estimated with qPCR.

Raw Illumina reads (92,365,800) were matched to the studied populations based on index read and then analyzed with *Stacks* software (Catchen et al. 2011; Catchen et al. 2013). First, reads were demultiplexed and cleaned with the *process_radtags.pl* program resulting in 71,611,749 reads for further analysis (Table S1.2). The SphI recognition site sequence (CATGC) was removed from all reads, since including this fragment could cause underestimation of total nucleotide diversity (Hohenlohe et al. 2010). Subsequently, for each individual, loci were reconstructed with the *denovo_map.pl* program including *MySQL* graphical visualization. For further analysis we used loci

present in all six populations and genotyped in at least 75% of the individuals of each population (see Supporting information S1.2).

Statistical analyses and environmental correlations

For all the analyses rove beetles sampled from different sites were assumed to represent local populations. Mitochondrial mtATP6 sequences were analyzed with DnaSP (Rozas 2009) and Arlequin 3.5 (Excoffier and Lischer 2010) aiming at estimation of basic population genetic statistics such as haplotype diversity, nucleotide diversity, number of polymorphic sites and measures of population differentiation. Pairwise genetic differences between studied populations were calculated based on haplotype frequency (F_{ST}) with 10,100 permutations. Median-Joining haplotype networks were constructed with Network 4.6 (Bandelt et al. 1999; http://www.fluxus-engineering.com) to illustrate genetic relationships among the haplotypes.

RADseq data were analyzed with the *populations* program of *Stacks*. Population genetic statistics including number of private alleles, number of haplotypes, haplotype diversity and nucleotide diversity were estimated. Pairwise differentiation between populations (F_{ST}) was estimated with Arlequin based on the allele frequency of SNPs (10,100 permutations). Number of effective migrants *Nm* per generation was estimated from $F_{ST} = 1/(1 + 4Nm)$ after Wright (1931). The extent of population structuring was also examined using a Bayesian clustering method implemented in *Structure* software (Pritchard et al. 2000; Falush et al. 2003, 2007; Hubisz et al. 2009). *Structure* does not use predefined assignment of individuals to populations and may be used to estimate the most likely number of genetically differentiated clusters and the fraction of the individual's genotype attributable to each cluster. Structure analyses were performed with the use of 100,000 burn-in steps and 500,000 post-burn-in iterations. All six populations were analyzed together, with one randomly chosen SNP per RAD tag (*--write_single_snp*), in total 615 loci. We tested K values range of 1 – 10 with 10 replicates for each value. Optimal K was selected with the Evanno method (Evanno et al. 2005) implemented in Structure Harvester (Earl and von Holdt 2012) and on the basis of the probability of the data given the number of clusters for various K values.

Significance of differences in haplotype and nucleotide diversity between populations was tested with a randomization test (1000 draws; mtDNA) and t-test (RADseq). Strict Bonferroni correction for multiple comparisons was applied. As metal concentrations in soil, both total and extractable, were highly correlated (Table S3.1), for further analyses Cd concentration was used as an index of soil pollution. Correlation of pollution level (total and 0.01 M $CaCl_2$-extractable concentrations of Cd), soil properties (pH, OM) and population genetic diversity was assessed with Pearson correlation analysis performed with SPSS 21 IBM Statistics.

To test the effect of pollution on the degree of genetic differentiation between populations (F_{ST}) while accounting for geographic distance, we performed a partial Mantel test (Smouse et al. 1986) with the use of IBDWS (Jensen et al. 2005; http://ibdws.sdsu.edu/). Isolation by distance was tested with a simple Mantel test. For both Mantel tests 10,000 randomizations were done. We used the following distance matrices: pairwise F_{ST} values, log-transformed geographic distance (straight line), difference in Cd total soil concentration.

To search for candidate loci under local selection we used two approaches. First, we used FDIST2 method (Beaumont and Nichols 1996) as implemented in Arlequin to identify SNPs that show extreme allele frequency differences between studied populations. FDIST generates the expected distribution of F_{ST} given expected heterozygosity under the island model and uses this distribution to identify the outlier loci. We performed 200,000 coalescent simulations. A locus was considered as outlier if its F_{ST} fell within the 1% tails of the simulated global F_{ST} distribution. Second, we used BAYENV software to search for loci whose allele frequency was correlated with pollution level (Coop et al. 2010). To describe the effect of soil contamination on allele frequency of each SNP, Bayes factors were generated with the use of Markov chain Monte Carlo algorithms (1,000,000 iterations). Soil contamination level was described as standardized Cd total and extractable concentrations. Standardization was performed by subtracting the mean and then dividing it through by the standard deviation of Cd concentration across populations.

Results

mtATP6

We identified 11 mitochondrial haplotypes among 96 sequenced *S. erythropterus* individuals (Table 2). The most common haplotype, observed in 75 individuals, was shared by all populations (Fig. 2). It differed from the remaining haplotypes by 1 - 6 mutations. The number of haplotypes ranged from only one in the OL7 population to seven in PK. The highest haplotype diversity was observed in the PK population ($H_d = 0.850 \pm 0.054$) characterized also by the highest nucleotide diversity ($\pi = 0.00300 \pm 0.00042$). Among the Olkusz populations, OL2 was the most diverse ($H_d = 0.450 \pm 0.151$; $\pi = 0.00242 \pm 0.00137$) while OL7 originating from the unpolluted site showed no mtATP6 diversity at all. None of the pairwise mtDNA F_{ST} indices among Olkusz populations were significant (Table 3). However, significant differentiation was observed between each Olkusz population compared separately with PK. When five Olkusz populations were considered as one panmictic population and compared with PK, mtDNA $F_{ST} = 0.4554$ was significant ($p < 0.001$). Genetic diversity ($H_d = 0.212 \pm 0.061$, $\pi = 0.00075 \pm 0.00034$) in the pooled Olkusz sample was significantly lower than in the PK population (randomization test; $p < 0.05$).

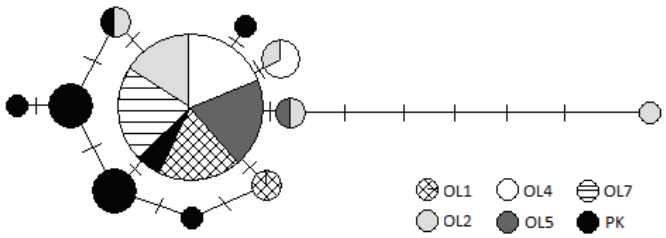


Fig. 2. Haplotype network of ATP6 sequences of *Staphylinus erythropterus* from metal-polluted and control sites, constructed using the Median-Joining method with Network 4.6. Circles represent distinct haplotypes. The size of each circle is proportional to the total number of individuals showing that haplotype and haplotype distribution over the populations is indicated with pie charts; the smallest circle corresponds to N=1. The genetic distance between haplotypes is measured by substitutions indicated by dashes along the connecting lines.

Table 2. MtDNA variation in *Staphylinus erythropterus* populations from metal-contaminated and clean sites in southern and central Poland. See Table 1 for further information on sampling sites and metal concentrations; Site – sampling location, N – number of analyzed individuals, N_{DP} – number of individuals with at least one double peak in electropherogram, Sex - number of females and males sampled, S – number of polymorphic sites, h – number of haplotypes, H_d – haplotype (gene) diversity (mean ± SD), π – nucleotide diversity per site (mean ± SD); computed with *DnaSP*; means with different letters are significantly different (randomization test; $p < 0.003$ after strict Bonferroni correction).

Site	N (N_{DP})	Sex	S	h	H_d	π
OL1	16 (4)	6♀ 10♂	1	2	0.233 ± 0.126 [a]	0.00051 ± 0.00027 [a]
OL2	16 (6)	9♀ 7♂	8	5	0.450 ± 0.151 [ac]	0.00242 ± 0.00137 [ac]
OL4	16 (2)	9♀ 7♂	1	2	0.233 ± 0.126 [a]	0.00051 ± 0.00027 [a]
OL5	16 (2)	10♀ 6♂	1	2	0.125 ± 0.106 [a]	0.00027 ± 0.00023 [a]
OL7	16 (6)	8♀ 8♂	0	1	0 [b]	0 [b]
PK	16 (1)	11♀ 5♂	5	7	0.850 ± 0.054 [c]	0.00300 ± 0.00042 [c]
Olkusz	*80 (20)*	42♀ 38♂	*9*	*6*	*0.212* ± 0.061	*0.00075 ± 0.00034*
all	*96 (21)*	53♀ 43♂	*12*	*11*	*0.388 ± 0.064*	*0.00143 ± 0.00034*

Table 3. Pairwise genetic differentiation between *Staphylinus erythropterus* populations from metal-contaminated and clean sites in southern Poland. See Table 1 for further information on sampling sites and metal concentrations: mtDNA F_{ST} based on haplotype frequency - below diagonal, RADseq F_{ST} - above diagonal. Significant values are marked with asterisks (10,100 permutations; $p < 0.003$ after strict Bonferroni correction).

	OL1	OL2	OL4	OL5	OL7	PK
OL1		0.0187*	0.0209*	0.0199*	0.0220*	0.0279*
OL2	0.0061		0.0109	0.0124	0.0176*	0.0220*
OL4	0.0044	-0.0170		0.0138	0.0167*	0.0219*
OL5	0.0029	0.0187	0.0029		0.0196*	0.0257*
OL7	0.0667	0.1000	0.0667	0.0000		0.0218*
PK	0.3067*	0.1961*	0.3067*	0.3633*	0.4333*	

RADseq

The filtering step was passed by 615 RAD loci containing 3017 SNPs. Although the largest number of haplotypes (h) and haplotype diversity (H_d) was found in the OL2 population (h = 3.26 ± 0.06, H_d = 0.324 ± 0.009; mean ± SE), the values were not significantly different from the other populations (Table 4). Population OL2 was also characterized by the highest number of unique variable positions (Private = 239). The level of genome-wide nucleotide diversity (π) was the same in all analyzed populations.

Pairwise F_{ST} ranged from 0.0109 to 0.0279 (Table 3). Global F_{ST} was equal to 0.0218. Differentiation between OL2, OL4 and OL5 populations was not significant. The estimated number of migrants *Nm* was equal to 11.2 per population per generation. No population structure was suggested by the Bayesian clustering. According to the value of the estimated probability of the data (lnP(D); Fig. S2 A) there was only one genetic group of rove beetles (K = 1). However Evanno method suggested K = 3 as the most likely number of genetic clusters (Fig. S2 B) but this method is not able to assess validity of K = 1.

Table 4. Genetic statistics of *Staphylinus erythropterus* populations from metal-contaminated and clean sites in southern and central Poland, estimated from RADseq data (final set of 615 RAD tags) for all nucleotide positions; N - genotyped individuals (mean), Bases - total number of analyzed nucleotide positions, Private - private variable positions, h - number of haplotypes (mean ± SE), H_d - haplotype (gene) diversity (mean ± SE), π - nucleotide diversity (mean ± SE); means with different letters are significantly different (t-test; $p < 0.003$ after strict Bonferroni correction).

Site	N	Bases	Private	h	H_d	π
OL1	15.2	54072	186	2.99 ± 0.06 [a]	0.315 ± 0.009 [a]	0.0047 ± 0.0002 [a]
OL2	15.4	54066	239	3.26 ± 0.06 [b]	0.324 ± 0.009 [a]	0.0048 ± 0.0002 [a]
OL4	15.3	54068	218	3.15 ± 0.06 [ab]	0.319 ± 0.009 [a]	0.0047 ± 0.0002 [a]
OL5	15.2	54065	200	3.12 ± 0.06 [ab]	0.317 ± 0.009 [a]	0.0047 ± 0.0002 [a]
OL7	15.3	54068	205	3.12 ± 0.06 [ab]	0.320 ± 0.009 [a]	0.0048 ± 0.0002 [a]
PK	14.7	54075	232	3.06 ± 0.06 [ab]	0.317 ± 0.009 [a]	0.0047 ± 0.0002 [a]

Environmental correlations

Neither haplotype diversity nor nucleotide diversity based on mtDNA or RADseq data correlated with cadmium concentrations in soil ($p > 0.05$). Also, no correlations of genetic diversity with soil pH and organic matter content were found (Table S3.2). The Mantel test (mtDNA F_{ST}: $Z = 3.95$, $r = 0.882$, $p = 0.003$; RADseq F_{ST}: $Z = 0.434$, $r = 0.750$, $p = 0.041$) showed significant isolation by distance. When PK populations were excluded from the analysis, this correlation was significant for mtDNA F_{ST} ($r = 0.780$, $p = 0.016$) but not for RADseq F_{ST} ($r = 0.454$, $p = 0.251$). As mtDNA F_{ST} within the Olkusz area was not significantly different from zero, isolation by distance between Olkusz sites should be interpreted with caution. Genetic differentiation was not correlated with pollution level when controlling for geographic distance (partial Mantel test; mtDNA: $r = -0.394$, $p = 0.973$; RADseq: $r = 0.329$, $p = 0.179$; Table S3.3, Fig. S3). In the FDIST scan for outliers, we found 18 loci (0.6%) possibly under diversifying selection at 0.01 significance level (Fig. S4). Sparse SNPs with $BF > 1$ were found by BAYENV (Fig. S5) and they were not consistent with FDIST outliers.

Discussion

According to the genetic erosion hypothesis (Van Straalen and Timmermans 2002) a decrease of population size due to exposure to metal pollution may lead to reduced genetic diversity at more polluted sites. We assessed genome-wide genetic diversity of *S. erythropterus* populations in a gradient of metal pollution but failed to detect an effect of pollution on the genetic diversity level. All analyzed populations showed very similar diversity at the nuclear genome, and observed nucleotide diversity ($\pi \sim 0.5\%$) was within the range reported for Arthropods by Leffler et al. (2012). Mitochondrial DNA diversity showed significant differences between some populations. However, none of the measures of genetic diversity in either genome was correlated with concentrations of metals in soil or with such soil properties as pH and organic matter content. Thus, our data taken at face value do not support the genetic erosion hypothesis although concentrations of metals observed at Olkusz transect were shown to have adverse effects on beetles (Stone et al. 2001; Skalski et al. 2010; Bednarska and Stachowicz 2013). We are aware that except pollution other factors like habitat fragmentation, small population size or low migration are necessary to co-occur with pollution to

cause genetic isolation and affect population genetic diversity. Species ecology and its history together with stochasticity determine population reaction to pollution. By studying additional species with different migratory potential we looked for circumstances under which genetic erosion hypothesis is valid.

Why did we not find an effect of long-term exposure to elevated metal concentrations on genetic diversity in *S. erythropterus*? Several methodological difficulties are discussed by Hoffmann and Willi (2008). The most likely explanation is a high level of gene flow among local populations as evidenced by little differentiation of allele frequencies between sites. In the Olkusz area no significant differentiation was observed in mtDNA while for nuclear RADseq markers some pairwise F_{ST} values were significant, but they were generally lower than 0.02. This indicates that although gene flow between *S. erythropterus* populations does not completely homogenize the populations, it is sufficient to prevent substantial differentiation of allele frequencies. At a larger geographic scale differentiation between populations from the Olkusz area and the Kozienicka Forest was more pronounced in mtDNA, although still low for nuclear markers ($F_{ST} < 0.03$). This result was corroborated by the Bayesian clustering approaches which, on the basis of nuclear variation, supports the presence of a single genetic cluster. The number of migrants (*Nm*, based on nuclear data) calculated under the island model assumption was 11.2 per generation. Although *Nm* estimation from F_{ST} should be interpreted with extreme caution (Whitlock and McCauley 1999), we provide these values just to illustrate that the migration rate is certainly much higher than one effective migrant per generation which is sufficient to prevent substantial differentiation due to the action of drift (Charlesworth and Charlesworth 2010). Intensive gene flow may mask the effect of pollution on genetic diversity in natural populations even for soil invertebrates of which many species seem to have low dispersal capacity (Costa et al. 2013). High dispersal was also suggested by Berckmoes et al. (2005) to explain the absence of an impact of metal contamination on microsatellite diversity in the wood mouse which is a relatively mobile species. Similarly, Lagisz et al. (2010) identified gene flow as a possible reason for difficulties in detecting effects of metal pollution on genetic diversity in populations of the ground beetle *P. oblongopunctatus*. Theodorakis et al. (2001) studying populations of the kangaroo rat (*Dipodomys merriami*) exposed to radionuclides also concluded that migration masked the genotoxic

effects of radiation. Considering our results and the above mentioned research, we believe that careful selection of study species is necessary to resolve the puzzle of pollution effects on genetic diversity in natural populations. Thus we decided to study the effect of pollution on genetic diversity of a range of species differing in their dispersal capabilities. Such a broad-scale study allows for evaluating the impact of both pollution and migration on genetic diversity in population chronically exposed to toxic chemicals. This article is the first one in the row of studies we have conducted.

A remarkable pattern was found in the population OL2 inhabiting the site with the highest availability of metals as proved by their 0.01 M $CaCl_2$-extractable concentrations (Table 1). Within the Olkusz area, animals from OL2 showed significantly higher mitochondrial diversity than other populations. At the nuclear level the number of haplotypes and haplotype diversity of OL2 were the highest, although not significantly different from those observed in other populations. Together with the largest number of private alleles (239 alleles; Table 4) these results may indicate a relatively larger effective population size at OL2 or the impact of metal pollution through increased mutation rates (Bickham et al. 2000). According to the neutral theory of molecular evolution, genetic diversity depends on new mutations that increase the level of diversity and genetic drift leading to the loss of genetic diversity. As genetic drift is slower in larger populations, populations with larger effective population size tend to have greater genetic diversity. Based on field observations and our sampling effort we suppose that the OL2 site indeed could be characterized by the highest number and density of *S. erythropterus*. Skalski et al. (2010) reported decreased beetle species abundance with increasing pollution level. However species abundant at all sampling sites (*Carabus arcensis, C. nemoralis, P. oblongopunctatus*) showed relatively higher abundance at site OL2, comparable with reference site. This may suggest larger population size of beetles inhabiting this site. Interestingly, density of earthworms, the main food source of beetles, was found to be the highest at polluted sites during Spring, but not other seasons (Tosza et al. 2010). We could suspect that high abundance of predators caused decrease of earthworms density at OL2 site. Perhaps this is particular pattern of Olkusz transect as Zvereva et al. (2010) in their meta-analysis reported significant adverse effects of pollution on population density of epigeic predators, including *Staphylinidae* and *Carabidae* beetles.

It is known that toxicity of heavy metals results in oxidative stress due to intensified generation of reactive oxygen species, ROS (Belyaeva et al. 2012). This can induce DNA mutations. Thus, one explanation for higher nucleotide and haplotype diversity at the OL2 site may be an increased mutation rate due to oxidative stress resulting from contamination. Migula et al. (2004) studying different beetle species, including *Staphylinus sp.*, along the same pollution gradient near Olkusz found correlations between some antioxidative enzymes activity and body concentrations of metals. Stone et al. (2002) observed the most elevated levels of enzyme activity (carboxylesterase , glutathione S-transferase) in ground beetle *P. oblongopunctatus* collected at polluted site OL2 (however in that paper this site is called OLK3). Increase of mutation rate due to pollution has already been mentioned by, for example, Eeva et al. (2006) who found the great tit *Parus major* population living near a copper smelter to have higher mtDNA nucleotide diversity than the population from an unpolluted area. Similarly, Štambuk et al. (2013) mentioned enhanced mutation load as possible explanation of the higher microsatellite diversity found in mussels *Mytilus galloprovincialis* from more polluted locations along the Adriatic coast. What is more, the whole Olkusz area was characterized by high number of individuals with double peaks in mtDNA sequence. If heteroplasmy was the source of these peaks, it would also suggest increased mutation rate. For example, Matson et al. (2006) reported an elevated level of heteroplasmy among marsh frogs at the polluted Sumgayit region and no heteroplasmy in the clean area.

On the other hand, dispersal influenced by pollution may lead to increased levels of genetic diversity at polluted sites due to immigration (Dallinger and Höckner 2013). Intensive immigration into the OL2 population could be another explanation of the higher genetic diversity observed at this site. It is possible that the OL2 site acted as an ecological sink. However, this is difficult to prove and we do not have definitive evidence supporting this hypothesis. It would be useful to sample more sites surrounding the smelter and assess direction of gene flow between numerous differently polluted sites.

What about local adaptation of *S. erythropterus* to metal pollution? Again, intensive gene flow may prevent populations from developing local adaptation (Lenormand 2002). Roughly, if migration rate exceeds the selection coefficient, no local adaptation is expected (Charlesworth and Charlesworth 2010). On the other hand, even if at one site there is local adaptation but at other sites there is no

selection against adapted alleles, these alleles will be spread over large areas by migrating individuals making selection undetectable from molecular data. Although we found some loci putatively under selection, their allele frequency was not correlated with contamination level. This suggests that pollution did not act here as a selection force. According to Rockman (2012) by performing genome scans we are able to detect polygenes of small-effect. Selection more often acts through such regions, not large-effect single genes. Overall, the hypothesis about local adaptation of *S. erythropterus* populations to metal pollution is not supported by our results. This is in agreement with the study of Lagisz and Laskowski (2008) who found no evidence of adaptation to metal pollution in ground beetle *P. oblongopunctatus* collected from the same Olkusz transect and reared at laboratory conditions till F_2-generation.

In conclusion, we analyzed genome-wide polymorphism data as recommended previously by, for example, Theodorakis et al. (2001) and Berckmoes et al. (2005). This kind of analysis should have maximum power to detect an effect of pollution on population genetic diversity. However, we did not find any significant effect in *S. erythropterus* populations. The very little genetic differentiation found among populations suggests that extensive gene flow among populations may erase potential effects of metal pollution.

Acknowledgements

We would like to thank Artur Kowalik and Ewelina Nowak from the Department of Molecular Diagnostics, Holycross Cancer Center, Kielce, Poland, for the possibility of using LabChip XT and technical training. We also thank Michał Stuglik for providing the necessary Python scripts. This study was supported by the Foundation for Polish Science International PhD Projects Programme co-financed by the EU European Regional Development Fund in the frame of the "Environmental Stress, Population Viability and Adaptation" project (MPD/2009-3/5) and the Polish National Science Center Grant no. 2011/03/N/NZ8/00013. Support from Jagiellonian University in Kraków, DS 758, is also acknowledged.

Appendix 4

Supplementary materials to Chapter 4

S1. Sequencing details

S1.1. Sequencing of mtATP6

Table S1.1. Information on mtATP6 sequence marker and PCR amplification: sequence length [bp], PCR primer sequences and PCR cycle scheme used for sequencing the mitochondrial gene of the rove beetle *Staphylinus erythropterus*.

Length [bp]*	Primer sequence	PCR profile
598 (465)	F1: TCTTTCGACCCATCTTCAGG	95^0C – 3 min
	R1: GGATAATTGCTACTGCTGATTCAA	30 cycles:
630	F2: TGATAACAAATTTATTTAGATCTTTCG	95^0C – 30 s
	R2: TGTTTGTGCATTAGTTTACTTCTCT	58^0C – 30 s
846	F3: TCGCCCTAATCCTTCTCATATT	72^0C – 1 min
	R3: AAGCTCCTAGTAGGGGTCAAGG	72^0C – 10 min

* - PCR product size including primer sequences; in brackets sequence length after trimming the ends

S1.2. Illumina sequencing and Stacks analysis

Rove beetle RAD tag library was mixed with another species RAD sample in 3:2 ratio and run on one HiSeq 2000 lane. The library comprised in total 16 different barcodes, which were used to distinguish individuals within a single population. Populations were distinguished based on unique 6 bp index sequence. To avoid problems due to low diversity of RAD library, sequencing was performed at relatively low cluster density (~650 K/mm^2) and with dedicated PhiX lane, but without sample PhiX spiking. In total sequencing resulted in 175,874,341 raw reads. These included 92,365,800 reads with rove beetle indices (exact matches only).

Raw Illumina reads were matched to the six studied populations based on the index read and then analyzed with *Stacks* software (Catchen et al. 2011; Catchen et al. 2013). First, reads were cleaned and demultiplexed with the *process_radtags.pl* program. We used only those reads having the correct barcode and high sequence quality. We applied sliding window quality filtering (*-w 0.15; -s 20*) and discarded reads failing Illumina chastity filter (*--filter_illumina*) as recommended by Zhou

and Rokas (2014). Any read containing an uncalled base was discarded (*-c*). As a decline in base quality score at the 3′-end is typical for Illumina sequence reads, the last three nucleotides were trimmed (*-t 93*). RAD tags with one sequencing error in the restriction enzyme overhang were rescued (*-r*). In total 22.5% of raw reads were discarded (Table S1.2).

Before further steps, we removed the SphI recognition site sequence (CATGC) from all reads. This resulted in final read length of 88 bp. Subsequently, for each individual, loci were reconstructed *de novo* requiring at least four identical reads to form a stack (-m 4). We allowed three nucleotides distance between stacks (*-M 3*) and no secondary reads (*-N M+0*). Highly repetitive RAD tags, marked as lumberjack stacks, were removed (*-t*). We set the maximum number of stacks allowed at a single locus to three (*--max_locus_stacks 3*) and enabled deleveraging algorithm (*-d*). A catalog across all individuals was constructed allowing two mismatches between catalog loci (*-n 2*). We decided to use the bounded-error implementation of the maximum-likelihood SNP calling model available in *Stacks* (Hohenlohe et al. 2010; Catchen et al. 2013). We calculated the error rate according to equation $\varepsilon = -(ln(1-x)/5)$ suggested by Emerson et al. (2010) resulting in a relatively low estimate of 0.012 for the analyzed dataset and we applied, in our opinion, rather conservative filtering for read quality. Hence, the upper bound of the error was reduced to the value of 0.025. The above mentioned steps were done with the *denovo_map.pl* program including *MySQL* graphical visualization. To avoid RAD tags from repetitive regions we used only those containing no more than eight SNPs (*export_sql.pl -F snps_u=8*). The *Stacks* parameters were chosen based on our knowledge about rather low polymorphism within *S. erythropterus* and a preliminary survey providing information about average coverage per RAD tag (Fig. S1.2). For further analysis we used loci present in all six populations and genotyped in at least 75% of the individuals of each population (*populations -p 6, -r 0.75*).

Table S1.2. Information on raw Illumina reads from HiSeq 2000 and reads filtered by *process_radtags* program for each sequenced rove beetle (*Staphylinus erythropterus*) population (OL1 – PK) and for all populations together (*ALL*). Populations were sampled at six field sites with different levels of metal pollution. See Table 1 for more information on site characteristics and metal concentrations.

Parameter/Pop	**OL1**	**OL2**	**OL4**	**OL5**	**OL7**	**PK**	***ALL***
index	CGATGT	TGACCA	GCCAAT	CAGATC	CTTGTA	ACAGTG	
raw reads	14,689,396	15,294,369	15,501,875	17,175,236	14,900,431	14,804,493	*92,365,800*
ambiguous barcodes	894,761 (6.1%)	897,659 (5.9%)	933,265 (6.0%)	1,075,432 (6.3%)	876,680 (5.9 %)	860,837 (5.8%)	*5,538,634*
failed chastity filter	1,728,146	1,879,350	2,011,745	2,086,550	1,819,083	1,758,898	*11,283,772*
ambiguous RAD tag	36,595	34,090	42,145	46,046	43,389	33,134	*235,399*
low QV reads	580,452	604,368	637,968	648,774	619,024	605,660	*3,696,246*
retained reads	11,449,442 (77.9%)	11,878,902 (77.7%)	11,876,752 (76.6%)	13,318,434 (77.5%)	11,542,255 (75.5%)	11,545,964 (78.0%)	*71,611,749 (77.5%)*

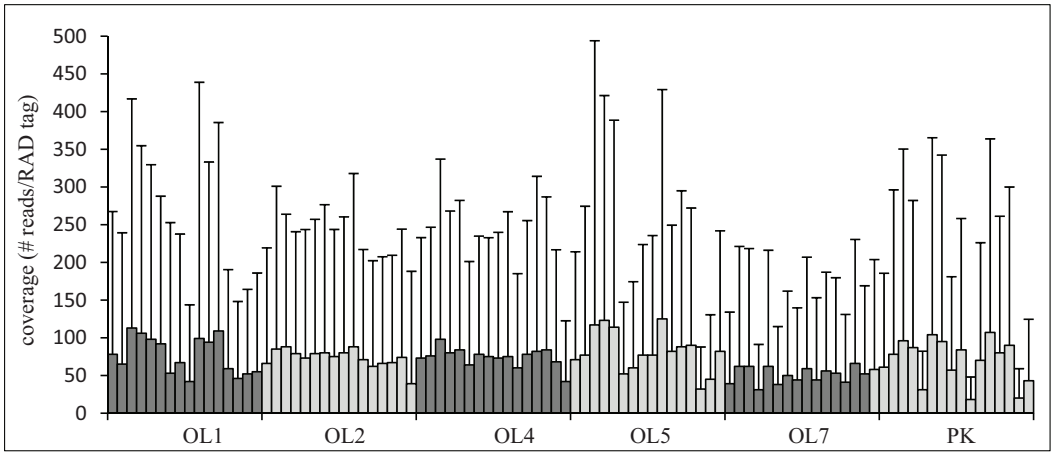
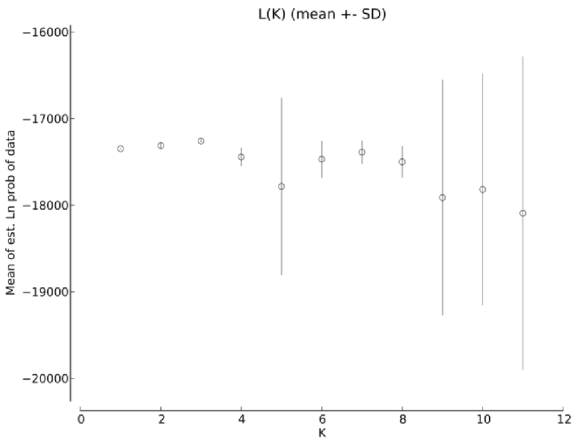
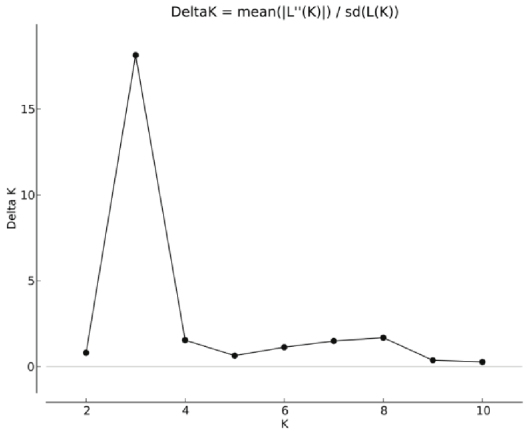


Fig. S1.2. Plot of coverage per RAD tag (mean ± SD) for individual rove beetles (*Staphylinus erythropterus*) samples at six field sites with different levels of metal pollution. See Table 1 for more information on site characteristics and metal concentrations. Data shown result from *denovo_map.pl* before merging stacks. Individual beetles are represented by single bars, grey tone separates populations.

S2. Genetic structure



A.



B.

Fig. S2. Estimation of K with Evanno method in Structure Harvester for *Staphylinus erythropterus* populations sampled along a metal pollution gradient and Structure results; A – L(K), B – ΔK; 100,000 burn-in steps, 500,000 replicates.

S3. Environmental correlations

Table S3.1. Pearson correlations of metals concentrations in soils at the six studied sampling sites; above diagonal – total concentration, below diagonal – 0.01 M $CaCl_2$-extractable concentrations; See Table 1 in main text for metal concentrations.

	Cd	**Pb**	**Zn**
Cd		r = 0.993, p < 0.001	r = 0.966, p = 0.002
Pb	r = 0.921, p = 0.009		r = 0.977, p = 0.001
Zn	r = 0.989, p < 0.001	r = 0.856, p = 0.029	

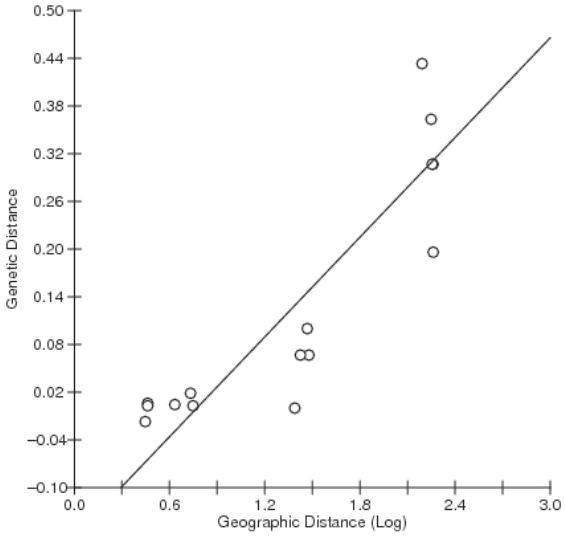
Table S3.2. Pearson correlations of genetic diversity measures (H_d and π) for *Staphylinus erythropterus* populations sampled at different field sites with metal concentrations (represented by Cd total and 0.01 M $CaCl_2$-extractable concentrations), soil pH and OM%.

	Cd_tot	**Cd_extr**	**pH**	**OM%**
H_d_mtDNA	r = -0.055, p = 0.918	r = 0.057, p = 0.915	r = 0.682, p = 0.135	r = 0.340, p = 0.510
H_d_RAD	r = 0.038, p = 0.943	r = 0.776, p = 0.070	r = 0.133, p = 0.802	r = 0.099, p = 0.851
π_mtDNA	r = 0.036, p = 0.946	r = 0.288, p = 0.579	r = -0.598, p = 0.210	r = 339, p = 0.510
π_RAD	r = 0.050, p = 0.925	r = 0.430, p = 0.395	r = 0.186, p = 0.724	r = -0.372, p = 0.467

Table S3.3. Mantel test and partial Mantel test statistics for *Staphylinus erythropterus* populations sampled at field sites with different levels of metal pollution; see Table 1 for detailed sites description.

	mtDNA Φ_{ST}	**RADseq F_{ST}**
Correlation of genetics and log (geographic distance)	Z = 3.95, r = 0.882, **p = 0.003**	Z = 0.434, r = 0.750, **p = 0.041**
Correlation of genetics and contamination (indicator) matrix	Z = 47.8, r = -0.190, p = 0.644	Z = 9.22, r = 0.214, p = 0.271
Partial corr. of genetics and log (geographic distance), controlling for indicator matrix	r = 0.898, **p = 0.007**	r = 0.768, **p = 0.038**
Partial corr. of genetics and indicator matrix, controlling for geography	r = -0.394, p = 0.973	r = 0.329, p = 0.179

A.



B.

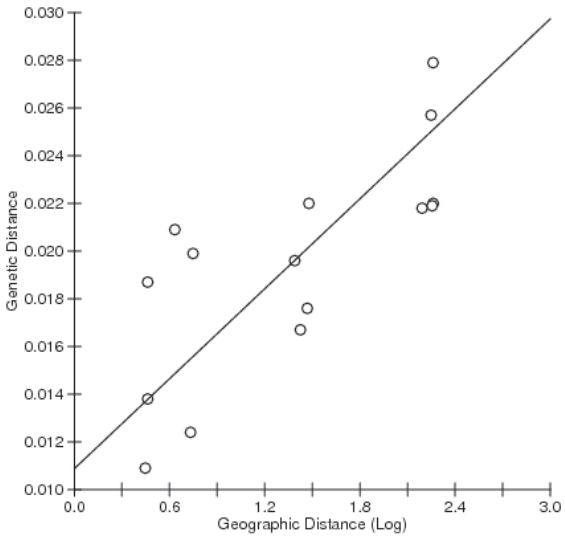


Fig. S3. Association of genetic distance and log(geographic distance) in *Staphylinus erythropterus* populations; A – mtDNA F_{ST} (r = 0.882; p = 0.003), B – RADseq F_{ST} (r = 0.750; p = 0.041); reduced major axis regression based on Mantel test.

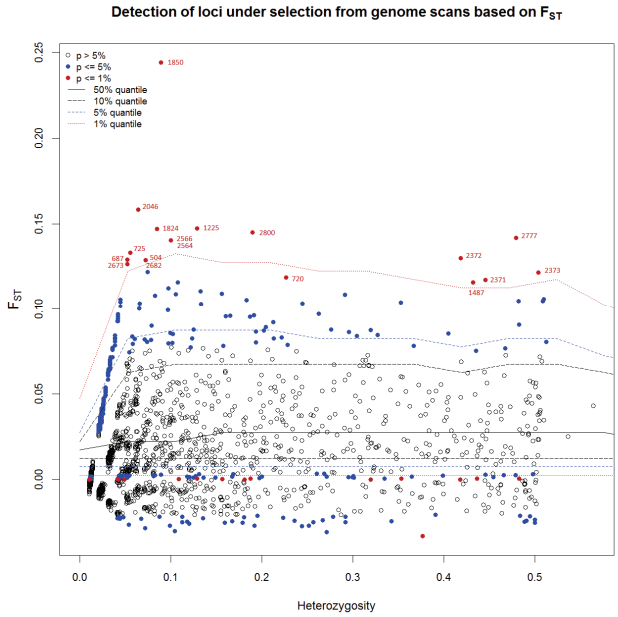


Fig. S4. Candidate outlier loci under directional selection detected in *Staphylinus erythropterus*. Detected 18 outliers at 0.01 significance level are marked with filled circles and their ID numbers; after 200,000 Arlequin simulations.

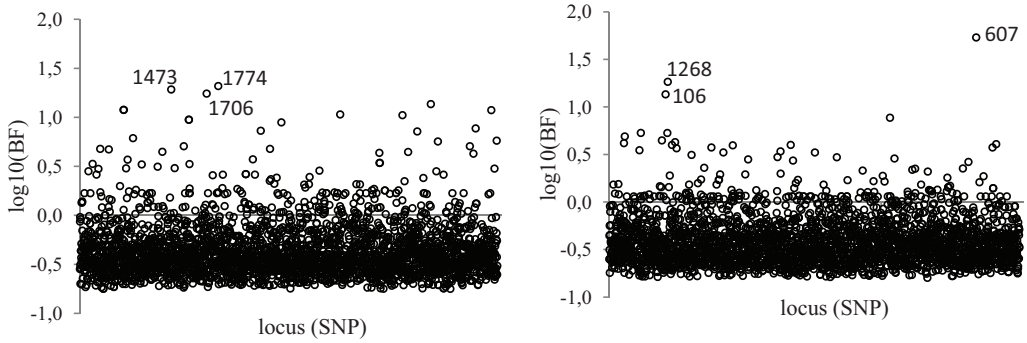


Fig. S5. Plot of Bayes factors ($log_{10}BF$) generated with BAYENV for each SNP in *Staphylinus erythropterus* for soil contamination effect: A - total Cd concentration, B - 0.01 M $CaCl_2$-extractable Cd concentration. Top hits loci are labelled with their ID numbers.